2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

A Partial Label-Based Machine Learning Approach For Cervical Whole-Slide Image Classification: The Winning TissueNet Solution*

Rutger H.J. Fick[1]**, Brice Tayart[1]**, Capucine Bertrand[1]**, Solène Chan Lang[1], Tina Rey[1], Francesco Ciompi[2], Cyprien Tilmant[3], Isabelle Farré[4] and Saima Ben Hadj[1]

Abstract— **Cervical cancer is the fourth most common cancer in women worldwide. To determine early treatment for patients, it is critical to accurately classify the cervical intraepithelial lesion status based on a microscopic biopsy. Lesion classification is a 4-class problem, with biopsies being designated as benign or increasingly malignant as class 1-3, with 3 being invasive cancer. Unfortunately, traditional biopsy analysis by a pathologist is time-consuming and subject to intra- and inter-observer variability. For this reason, it is of interest to develop automatic analysis pipelines to classify lesion status directly from a digitalized whole slide image (WSI). The recent TissueNet Challenge was organized to find the best automatic detection pipeline for this task, using a dataset of 1015 annotated WSI slides. In this work, we present our winning end-to-end solution for cervical slide classification composed of a two-step classification model: First, we classify individual slide patches using an ensemble CNN, followed by an SVM-based slide classification using statistical features of the aggregated patch-level predictions. Importantly, we present the key innovation of our approach, which is a novel partial label-based loss function that allows us to supplement the supervised WSI patch annotations with weakly supervised patches based on the WSI class. This led to us not requiring additional expert tissue annotation, while still reaching the winning score of 94.7%. Our approach is a step towards the clinical inclusion of automatic pipelines for cervical cancer treatment planning.**

Clinical relevance— **The explanation of the winning TissueNet AI algorithm for automated cervical cancer classification, which may provide insights for the next generation of computer assisted tools in digital pathology.**

I. Introduction

Cervical cancer is the fourth most common cancer in women worldwide [6]. The reduction of cervical cancer mortality is possible with earlier treatment through an early diagnosis at the precancer stage. To make this diagnosis, a tissue sample from the uterine cervix is extracted, fixed in a paraffin glass slide, stained with hematoxylin and eosin (H&E) and examined at a high resolution by an experienced pathologist with a microscope. However, glass slides diagnosis through visual inspection is time-consuming and subject to inter- and intra-pathologist variation. In the hope to reduce this variability and diagnosis time, glass slides can be digitalized to create ultra-high resolution Whole Slide Images (WSI), which enables the development of image processing algorithms to assist the pathologist and standardize diagnosis.

In this context, the TissueNet data challenge [9] was organized to allow machine learning approaches to compete for the automatic detection of the epithelial lesions of the uterine cervix. Lesion classification is a 4-class problem, with biopsies being designated as benign or increasingly malignant as class 1-3: LSILs (class 1) is Low-grade squamous intraepithelial lesion confined to the inner one-third of the epithelium. HSILs (class 2) is High-grade squamous intraepithelial lesion, spread beyond the lower third of the epithelium. SCC (class 3) is squamous cell carcinoma corresponding to an invasive cancer [2].

The challenge featured one thousand expert-labeled Whole Slide Images (WSIs) collected from medical centers across France. Annotations are equisized squares focused in and around the epithelium: the area of interest for cervical cancer diagnosis. Therefore, most machine- and deep learning-based approaches [see comprehensive review 5] first segment the epithelium from non-epithelium tissue, after which epithelium-only tissue features are used to predict the slide label. The TissueNet data and annotations, however, do not lend themselves easily to training an epithelium segmentation algorithm as only square annotations with a single label were provided. To succeed, contestants needed to overcome this limitation, besides the usual problems faced in histopathological image analysis such as stain, cell, and tissue variability.

Given these conditions, our winning end-to-end WSI classification approach is described in Figure 1 and consists of two main steps. First, we use a multi-resolution DenseNet ensemble [4] to predict the lesion class at patch level for all tissue in a slide, ignoring the distinction between epithelium and non-epithelium tissue. The slide label is then predicted with an SVM classifier trained using the statistical features calculated from aggregating patch-level predictions. We found that the lack of non-epithelium annotations causes many benign non-epithelium patches to be erroneously classified as cancer. Based on the knowledge that a slide cannot contain annotations more severe than its label, we supplement the TissueNet annotations by adapting the Hard Negative Mining (HNM) approach [7] to recover weakly supervised, *partially labeled* patches from benign, class 1 and 2 labeled WSIs [1]. We introduce a novel partial label-based loss to train on both labeled and partially labeled patches, which was key to achieve the best performance in the challenge.

The article is organized as follows: In Section II, we detail

*This work has been performed on data that were available during the 2020 Data Challenge of the French Society of Pathology and the Health Data Hub, with the support of Grand Defi for A.I. in Health.

** These authors contributed equally to this work.

[1]Tribvn Healthcare, Paris, France; [2]Radboud UMC, Nijmegen, The Netherlands; [3]GHICL, Lille, France; [4]XPath Nord, Leulinghem, France.

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

each step of our analysis pipeline and present the partial labeling-based loss function. In Section III, we present our challenge results and show the quantitative and qualitative effects of including partially labeled patches. Finally, in Section IV, we discuss the choices made and possible improvements.

II. Materials and Methods

In this section we present the TissueNet data, our novel partial label loss, and our end-to-end processing pipeline.

A. TissueNet Data

The data consists of 1015 annotated WSIs for supervised training. In total, there were 5926 local square annotations within these labeled slides, consisting of 300μm×300μm squares indicating graded tissue areas: 0) benign (normal or subnormal), 1) low malignant potential, 2) high malignant potential, and 3) invasive cancer. Annotations are found around the epithelium region, equally distributed between the four classes. The WSI images are acquired at 0.25μm/pix resolution (level 0), but are structured into pyramidal "levels" to enable loading at lower resolutions by factors of 2, such that levels [1, 2, ...] are [0.5, 1.0, ...]μm/pix.

B. Defining A Partial-Label Cross-Entropy Loss for Weakly Supervised Patches

Partially labeled, multi-class classification is a problem where instead of a single label per instance, the algorithm is given a candidate set of labels, only one of which is correct [1]. Applied to cervical cancer classification, we know that patches cannot be classified as a higher class than the slide they originate from. This means a benign slide cannot contain class 1-3 patches, a class 1 slide cannot contain class 2-3 patches, and a class 2 slide cannot contain class 3 patches. Patches thus inherit a weak partial label from the slide label, where we know it *cannot* be any higher class than the slide, but have no information on which of the valid candidates is the true class.

Let $\mathbf{x}^{(m)}$ be the sample image and $y^{(m)}$ the image class for sample m. In the case of partial labeling, $y^{(m)}$ is not known but instead a set $G^{(m)}$ is known such that $y^{(m)} \in G^{(m)}$. The goal is then to find a loss function that still allows a classifier to learn from these instances despite the ambiguous labeling.

Previously proposed loss functions include partially labeled binary cross-entropy (BCE) applied to multi-class classification [1]. However, we found that using this straightforward approach actually decreased performance: false negatives increased faster than false positives decreased.

Instead, we chose to develop a categorical cross-entropy (CCE) loss adapted to partially labeled data, which as far as we know was not previously used in the literature. This approach has the advantage of being interpretable in terms of probabilities (compared to sigmoid-based multi-class BCE) and is a generalization of the standard CCE, which we used for labeled patches. We define a loss that minimizes the difference between the (softmax) output of a network $\hat{\mathbf{y}}$ and partially labeled pseudo-label $\bar{\mathbf{y}}$, which we define below. We denote $\mathbf{z}^{(m)} = g(\mathbf{x}^{(m)}) \in \mathbb{R}^L$ the vector of logits of length the number of class predictions L given by network g. Dropping the image indexing (m) for convenience, we define $\hat{\mathbf{y}}$ the network prediction after a softmax layer as $\hat{\mathbf{y}} = [\hat{y}_0, \hat{y}_1, \hat{y}_2, \hat{y}_3]$, where

$$\hat{y}_i = \frac{\exp(z_i)}{\sum_{k=0}^{L-1} \exp(z_k)}, \quad i :\in 0 \dots L-1. \tag{1}$$

The categorical cross entropy is then computed as

$$\mathcal{L}_{CCE}(\hat{\mathbf{y}}, \bar{\mathbf{y}}) = \sum_{i=0}^{L-1} -\bar{y}_i \log \hat{y}_i$$

.

We design the pseudo-label $\bar{\mathbf{y}}$ to enforce four conditions:

1) Normalization to one: $\sum_{i=0}^{L-1} \bar{y}_i = 1$;
2) Zero labeling of non-candidate classes: $\bar{y}_i = 0, \forall i \notin G$;
3) The labeling should not modify a valid solution. If $\hat{y}_i = 0, \forall i \notin G$, then $\bar{\mathbf{y}} = \hat{\mathbf{y}}$ and thus $\frac{\partial \mathcal{L}_{CCE}(\hat{\mathbf{y}}, \bar{\mathbf{y}})}{\partial \mathbf{z}} = 0$;
4) Finally, gradient neutrality towards in-set classes: $\frac{\partial \mathcal{L}_{CCE}(\hat{\mathbf{y}}, \bar{\mathbf{y}})}{\partial z_i} = \frac{\partial \mathcal{L}_{CCE}(\hat{\mathbf{y}}, \bar{\mathbf{y}})}{\partial z_j}, \forall i, j \in G$.

The neutrality condition is key to learning from partial labels using CCE. For example, in the case $\hat{\mathbf{y}} = [0.4, 0.1, 0.25, 0.25]$ and $G = \{0, 1\}$, we want to suppress the two invalid classes but also *equally* push the solution towards each class in G, since we have no information to prefer one over the other.

We will show that the following pseudo-label expression satisfies all four conditions:

$$\bar{y}_i = \begin{cases} \hat{y}_i + \frac{1}{|G|} \sum_{k \notin G} \hat{y}_k & , i \in G \\ 0 & , i \notin G. \end{cases} \tag{2}$$

Note that if $|G| = 1$ we return to the usual CCE one-hot encoded target label. For the example given this definition results in the pseudo-label $\bar{\mathbf{y}} = [0.65, 0.35, 0, 0]$, satisfying conditions 1) and 2). We now derive the CCE gradient using this pseudo-label to show it also satisfies gradient conditions 3) and 4).

We define the gradient by first defining the partial gradient for the CCE and softmax, with δ being the Kronecker symbol:

$$\frac{\partial \mathcal{L}_{CCE}(\hat{\mathbf{y}}, \bar{\mathbf{y}})}{\partial \hat{y}_i} = -\frac{\bar{y}_i}{\hat{y}_i} \qquad \frac{\partial \hat{y}_j}{\partial z_i} = \hat{y}_i(\delta_{ij} - \hat{y}_j) \tag{3}$$

where also $j \in 0, \dots, L-1$. The gradient with respect to $\mathbf{z}$ is then given by

$$\frac{\partial \mathcal{L}_{CCE}(\hat{\mathbf{y}}, \bar{\mathbf{y}})}{\partial z_i} = \sum_{j=1}^{L} \frac{\partial \mathcal{L}_{CCE}}{\partial \hat{y}_j} \frac{\partial \hat{y}_j}{\partial z_i} \tag{4}$$

$$= \sum_{j=1}^{L} \left(-\bar{y}_j \frac{\hat{y}_i(\delta_{ij} - \hat{y}_j)}{\hat{y}_j} \right) \tag{5}$$

$$= -\bar{y}_i + \sum_{j=1}^{L} \bar{y}_j \hat{y}_i \tag{6}$$

$$= \hat{y}_i - \bar{y}_i \tag{7}$$

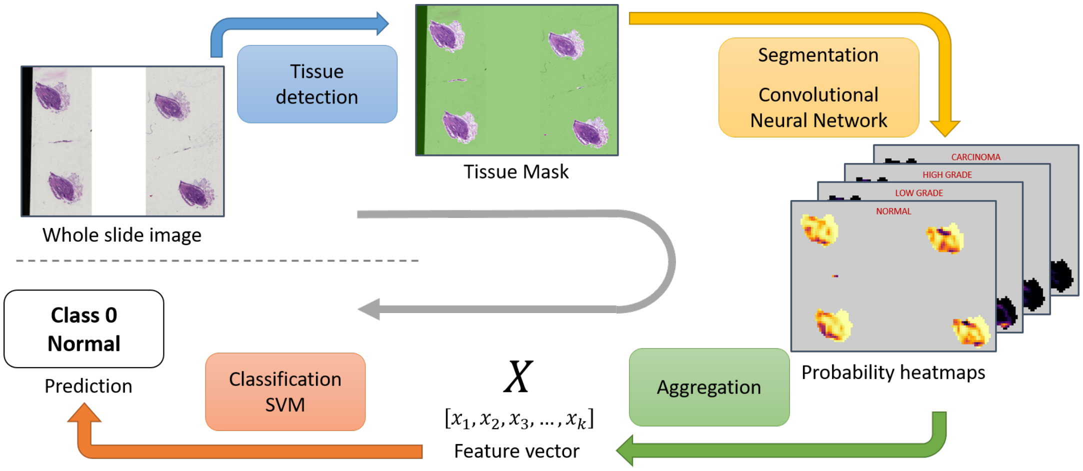

Fig. 1. Our end-to-end cervical cancer classification pipeline. We start by pre-processing the raw WSI to normalize and segment the foreground tissue mask. Then, we use a multi-resolution ensemble CNN to predict at patch level the probability of each of the four lesion classes. The ensemble was trained using both strongly supervised and weakly supervised, partially labeled patches. Next, we reduce the patch-level probabilities to a slide-level statistics feature vector. Finally, we feed these statistics into an SVM classifier to predict the slide-level lesion status.

where simplification (7) is due to condition 1: $\sum_{i=0}^{L-1} \bar{y}_i = 1$. Filling in Eq. (2) for $\bar{y}$ results in

$$\frac{\partial \mathcal{L}_{CCE}(\hat{\mathbf{y}}, \bar{\mathbf{y}})}{\partial z_i} = \begin{cases} -\frac{1}{|G|} \sum_{k \notin G} \hat{y}_k & , i \in G \\ \hat{y}_i & , i \notin G. \end{cases} \tag{8}$$

We verify that indeed the gradient is zero everywhere when $\sum_{k \notin G} \hat{y}_k^{(i)} = 0$, satisfying condition 3). When this is not the case, we see that the gradient is equal for all classes in the valid set, satisfying condition 4). Indeed, filling in the example we gave before, the gradient would be $[-0.25, -0.25, 0.25, 0.25]$, pushing the solution equally towards classes 0 and 1, and away from invalid classes 2 and 3 by their softmax value.

C. End-to-End TissueNet Processing Pipeline

Our end-to-end processing pipeline consists of 4 steps, as shown in Figure 1.

1) Tissue Detection: We normalize the brightness of the input WSI and segment the foreground tissue mask at image level 6 (16 μm/pix), which offers a compromise between the execution time and the segmentation quality. Brightness normalization ensures the WSI background to be white (no absorption). We guarantee this by selecting a frame close to the outer edge of the scanned area - where no tissue is expected - and take the median of each channel and divide the image by this value. After normalization, we use Otsu's method to detect tissue areas, followed by some morphological filtering opening/closing to clean small false positive and holes in the mask.

2) Multi-Resolution, Ensemble-Based Patch-wise Classification: We use a multi-resolution ensemble CNN to predict benign and 1-3 class probabilities per patch.

a) Patch Resolution: Accurate lesion classification requires both sufficiently high resolution and enough context on the full thickness of the epithelium. Therefore, our ensemble CNN uses a range of patch sizes and resolutions: 256×256 at level 3 (2μm/pix, high resolution, less context), 256×256 at level 4 (4μm/pix, lower resolution, more context) as well as 384×384 at 2μm/pix.

b) CNN Ensemble: For our CNN ensemble we use DenseNets [4], as we found it outperformed ensemble compositions using ResNets. After experimentation, We found the best performing ensemble used DenseNet169 on level 4 patches and DenseNet121 on Level 3 patches, both pretrained on ImageNet. Having tried several combinations, we chose a linear combination of the five models (two on level 4, three on level 3), followed by a softmax layer to output the final four class probabilities per patch, see Table I.

c) Data Augmentation and Training Parameters: We augment our data using digital pathology specific HED color augmentation [8], random linear transformation, Cutout [3] and CutMix [10]. To train each CNN, we use 200 epochs with a learning rate of 5×10^{-2} and use Glorot weight initialization. For classification, we use the categorical cross-entropy loss and add a ℓ_2 weight regularization with $\lambda \in [0.01 - 0.5]$ depending on the dataset.

d) Hard Partial Label Mining: Annotations were given primarily in the epithelium, giving the network a cancer bias towards non-epithelium benign tissue. We use HNM [7] adapted to partially labeled data to enrich our training data. After one training round on annotated patches, we do a full inference on all benign, class 1, and class 2 tissue slides. For normal slides, we add misclassified lesion class 1, 2 and 3 patches to the training data. For class 1 slides we add misclassified class 2 and 3 patches and misclassified class 3 patches for class 2 slides. In the second round, these *partially labeled* patches are added to the annotated patches with a specialized partial-label loss, which we presented in Section II-B.

TABLE I

REPORTED ACCURACY OF THE SEPARATE AND PROPOSED MULTI-RESOLUTION ENSEMBLE NETWORKS. HERE PL REPORTS THE ACCURACY ONLY PARTIALLY LABELED PATCHES

Model	Patch Size	Resolution	Accuracy	Accuracy (PL)
DenseNet169	256× 256	4 µm/pix	79%	95%
DenseNet169	256× 256	4 µm/pix	79%	96%
DenseNet121	384× 384	2 µm/pix	80%	97%
DenseNet121	384× 384	2 µm/pix	81%	96%
DenseNet121	256× 256	2 µm/pix	81%	95%
ensemble			**85.0%**	**98.0%**

3) Slide-wise Patch Feature Aggregation: Depending on the tissue in the slides, the number of patches varies from less than 5 to over 7000. Patches were chosen to overlap 75%, averaging the predicted class probabilities in overlapping areas. The aggregation step consists in calculating, *for each class separately*, statistics on the previously calculated patch predictions, and concatenating them into a vector of fixed size. Specifically, we obtain a feature vector with histogram percentiles $\hat{\mathbf{X}} = [p_{80,i}, p_{90,i}, p_{95,i}, p_{99,i}]$ for lesion class index $i = 0, 1, 2, 3$.

4) Slide-wise SVM Classification: Given feature vector $\hat{\mathbf{X}}$, we use SVM to predict the probability distribution $\mathbf{p} = [p_0, \ldots, p_3]$ that the slide belongs to class [0-3]. We trained the SVM with linear kernel, regularization parameter $C = 0.5$ and kernel coefficient $\gamma = 0.21$, using 10-fold cross-validation on the patch prediction validation data. We also use the contest reward matrix $\mathbf{R} \in \mathbb{R}^{4\times4}$ to maximize the expected score. Our final prediction is chosen from gain $\mathbf{r} = \mathbf{pR}$ as $i^* = \text{argmax}_i \mathbf{r}_i$.

III. RESULTS

In this section, we present the results of this paper. In section III-A we present the effect of adding patches to the problem using the partial loss based on HNM. Then, in Section III-B we present our winning results for the TissueNet challenge.

A. Effect of Adding Partially Labeled Patches to Training

To show the effect of adding partially labeled patches to model training we use only the DenseNet121 working on patches of size 256 × 256 at 2μm/pix. In Figure 2 we show two different scenarios: using only pathologist-annotated patches (left) versus also including partially labeled patches (right). For each scenario, we show the patch-wise prediction confusion matrix on the validation set of TissueNet (top row), as well as an overlay of the patch prediction over a benign, class 1, 2 and 3 slide (bottom 4 rows). It can be seen that only using pathologist annotations results in patches often being classified as a higher lesion class than the slide label, which is inconsistent. Adding partially labeled patches to model training results in slides rarely containing patches with a higher lesion class than the slide label, improving subsequent slide label prediction. We therefore used this approach to create ensemble CNNs for challenge submission.

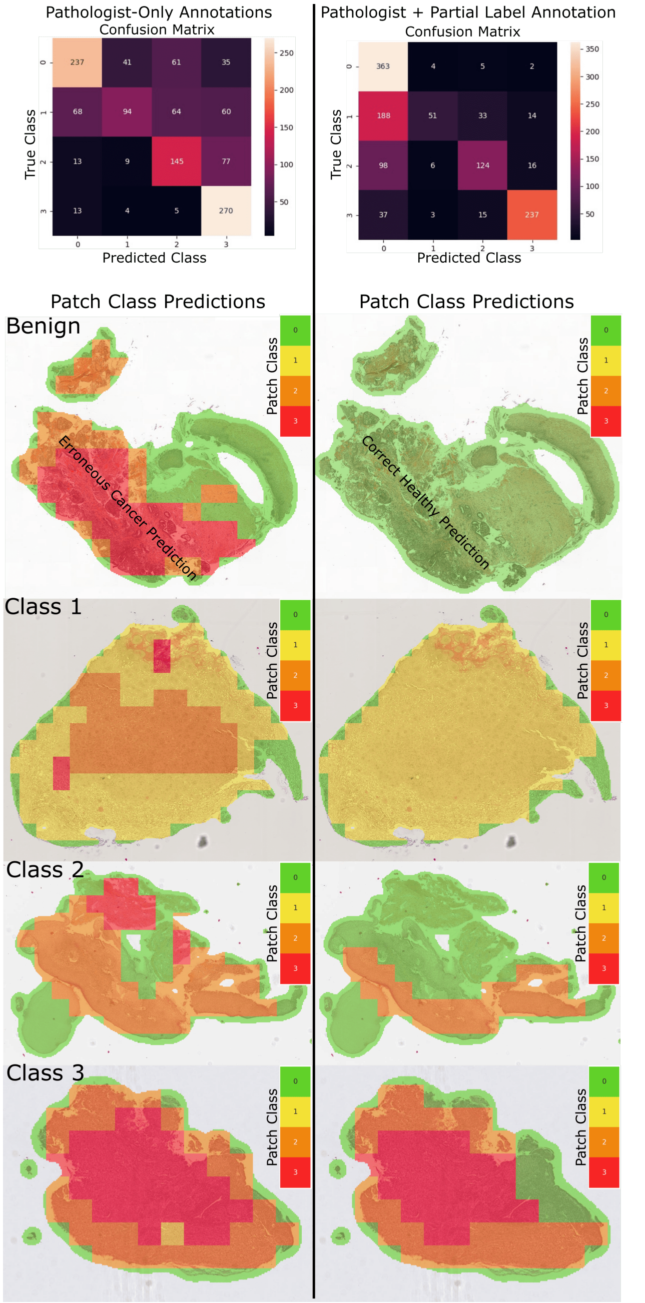

Fig. 2. The effect of including partially labeled patches in the training process, represented in confusion matrices (top row) and with patch classification overlay on a benign slide and class 1-3 slides (bottom 4 rows).

B. TissueNet Challenge Results

In Table I we evaluate the patch-wise prediction accuracy of separate DenseNet models at different resolutions, along with its ensemble. Notice that each separate model reaches an accuracy around 80% on multi-class accuracy, and over 95% on the partially labeled patches. The ensemble of the 5 models increases the multi-class accuracy to 85% and also slightly increases the partially labeled accuracy to 98%.

In Table II, on the top we show the confusion matrix of our patch-wise lesion class prediction, and on the bottom the confusion matrix of the subsequent WSI-level lesion class

TABLE II

The confusion matrices of our proposed patch-wise multi-resolution ensemble network (top) and our Aggregate-based WSI prediction using SVM (bottom). Note that errors are usually made between neighboring classes, but not disjoint classes. Top results on the validation set, bottom on the (now unavailable) test set of TissueNet.

Ensemble CNN Patch-wise Pred.

Class	Benign	class 1	class 2	class 3
Benign	**0.965**	0.024	0.010	0.0
class 1	0.252	**0.645**	0.096	0.007
class 2	0.075	0.130	**0.770**	0.025
class 3	0.010	0.0	0.024	**0.966**

Aggregate SVM WSI Pred.

Class	Benign	class 1	class 2	class 3
Benign	**0.764**	0.217	0.019	0.0
class 1	0.211	**0.690**	0.099	0.0
class 2	0.021	0.261	**0.685**	0.034
class 3	0.0	0.014	0.033	**0.953**

prediction after aggregation. In both cases most errors are made between adjacent classes. This is consistent with the pathologists' intuition, where differences between adjacent classes can be subtle. In both settings, lesions tend to be underclassified, with the most frequent error being class 1 being classified as benign tissue, and class 2 being classified as class 1. However, class 3 (cancer) is accurately classified.

IV. Discussion and Conclusion

In this work, we presented our winning end-to-end pipeline for whole slide image-based cervical cancer classification. The key points for our success were:

- The choice of context and the resolution (image-level) of the analysis, which was reached in consultation with expert pathologists;
- Our use of hard partial label mining and novel partial label CCE loss to counteract the annotation bias of benign tissue, which were mostly annotated in the epithelium and not the cervical stroma (Figure 2);
- Our use of multi-resolution ensemble models whose parameters are driven by different contexts, whose improvement we show in Table I;
- Taking into account the challenge reward matrix as a penalization matrix in the final class prediction.

Finding the best patch-prediction ensemble CNN was one of the hardest optimization problems in the challenge. The reason being that finding the best CNN composition is hard to measure, given that the ensemble's purpose is to generate patch probability features *from an entire slide* as input to an SVM, which predicts the final slide label. It was computationally infeasible to evaluate each trained model on the entire validation set of 203 slides through this pipeline, which can take up an hour for each set of possible hyperparameters. Instead, we chose ensemble composition based on a model's accuracy on only patches of the validation and normal slides. However, choosing the best model based on accuracy can also be ambiguous given that we estimate accuracy both on the pathologist annotated patches and the additional partially labeled patches.

Regardless of the model composition, it is still difficult to classify differences between adjacent classes, as there are many borderline cases, in which also the "ground truth" slide label will have large inter-observer variability. Regardless, we show in Table II that we make few "expensive" mistakes (classifying cancer as benign or vice-versa), and errors mainly exist in neighboring classes.

Finally, our approach led us to the winning score of 94.7%. While still not perfect, our approach is a step towards the clinical inclusion of automatic pipelines for cervical cancer treatment planning.

References

[1] Timothee Cour, Ben Sapp, and Ben Taskar. "Learning from partial labels". In: *The Journal of Machine Learning Research* 12 (2011), pp. 1501–1536.

[2] Teresa M Darragh et al. "The lower anogenital squamous terminology standardization project for HPV-associated lesions: background and consensus recommendations from the College of American Pathologists and the American Society for Colposcopy and Cervical Pathology". In: *Archives of pathology & laboratory medicine* 136.10 (2012), pp. 1266–1297.

[3] Terrance DeVries and Graham W Taylor. "Improved regularization of convolutional neural networks with cutout". In: *arXiv preprint arXiv:1708.04552* (2017).

[4] Gao Huang et al. "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[5] Chen Li et al. "A review for cervical histopathology image analysis using machine vision approaches". In: *Artificial Intelligence Review* 53.7 (2020), pp. 4821–4862.

[6] Hyuna Sung et al. "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". In: *CA Cancer J Clin.* (2021).

[7] Kah-Kay Sung. "Learning and example selection for object and pattern detection". In: *AI Technical Reports (1964 - 2004)* (1996).

[8] David Tellez et al. "Whole-slide mitosis detection in H&E breast histology using PHH3 as a reference to train distilled stain-invariant convolutional networks". In: *IEEE transactions on medical imaging* 37.9 (2018), pp. 2126–2136.

[9] *TissueNet: Detect Lesions in Cervical Biopsies*. URL: `https://www.drivendata.org/competitions/67/competition-cervical-biopsy/` (visited on 04/30/2021).

[10] Sangdoo Yun et al. "Cutmix: Regularization strategy to train strong classifiers with localizable features". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 6023–6032.